Comments on Woodward, "Making Things Happen"

Review of Woodward, Making Things Happen

Ned Hall

I have one complaint, which is that Woodward's subtitle misleads: "a theory of causal explanation" names just one part of what this excellent book contains. You will also find detailed, illuminating discussions of causation, laws of nature, events, the theory of confirmation, and the nature of good philosophical methodology, to name just the most prominent topics. The level of philosophical acuity is uniformly high, and Woodward's command of and ability to engage with a wide range of both philosophical and scientific material is extremely impressive. A must-read, then—and not just for philosophers of science and metaphysicians. Biased though I am, I still do not think it is an exaggeration to say that the quality of science education at the college level would improve significantly if every statistics course incorporated this book as required reading (or at least chapters 1 - 3, and chapter 7).

Here is a sketch of the ideas that lie at the heart of Woodward's approach to causation and explanation. Suppose you want to understand some system—an economy, or a machine, or a living organism, or a chemical reaction, what have you. Then scientific inquiry will give you such understanding by showing you how to describe the interacting parts of the system by means of variables whose values represent different possible states of those parts, and by means of "structural equations" which capture the relations of immediate dependency between these variables. Here are some illustrative examples, of a kind quite familiar from the contemporary causation literature. The following two illustrations depict distinct systems of interacting "neurons". These neurons are much simpler than the real thing: all they can do is either fire or not fire, and the connections between them come in just two flavors, stimulatory (represented by a normal arrow), or inhibitory (represented by a line with a blob at the end). Shading indicates that the neuron fires, and the order of events is left-to-right.

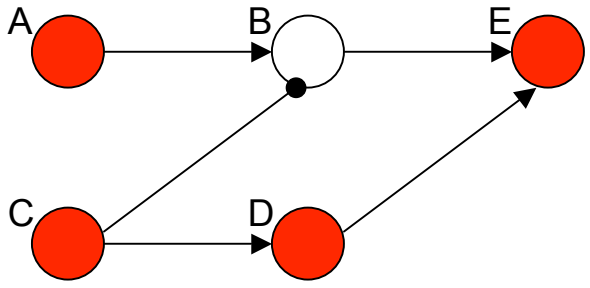

Figure 1

Thus in figure 1, neurons A and C fire simultaneously (at time 0, say). At time 1, neuron D fires as a result of the stimulatory signal from C; at time 2, E fires as a result of the signal from D. B, although stimulated by A, also receives an inhibitory signal from C, and so does not fire.

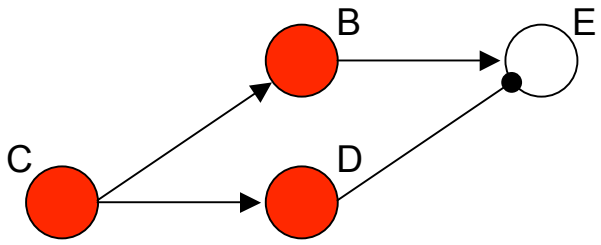

Figure 2

Similarly, in figure 2 neuron C fires, stimulating both B and D to fire. But the inhibitory signal from D cancels the stimulatory signal from B, so E does not fire.

Choosing variables and structural equations for these systems is trivial. In each case, we assign a binary variable to each neuron that takes the value one if that neuron fires and zero if it remains dormant. For figure 1, we have these equations:

E = B + D – BD
D = C
B = A(1 – C)

And for figure 2, we have these equations:

E = B(1 – D)
D = C
B = C

We will draw some specific lessons from these examples later on. For now, we can use them to illustrate some general features of Woodward's scheme. First, it is not the business of structural equations to describe merely de facto relationships between the values of variables (in a given kind of system, as it might be). Rather, they are intended to capture a certain kind of modal structure that the system manifests, consisting, roughly, in the facts about what the values of certain variables would have been if others had been set to certain values by a certain kind of "intervention". Thus, the second equation for figure 1 tells us that the value for variable **D** counter-

factually depends on the value of **C**, and **C** alone. This dependence is asymmetric—**C** does not likewise depend on **D**—and is to be understood in such a way as to rule out "backtracking"—so that we are not allowed to say that if the value of **B** had been different, then that would have been because the value of **C** was different, and so the value of **D** would have been different as well. Woodward's technical notion of "intervention" is designed to help make precise the particular kind of counterfactual dependence that is at issue here. (It is also, it is worth noting, a pretty good regimentation of the ordinary notion of intervention.)

It's worth pausing to see how, *given* a certain system of structural equations, we can provide truth-conditions for a certain kind of conditional construction. Specifically, if we wish to evaluate

If X = v, then P

where **X** is some variable, **v** some possible value for it, and **P** some claim whose truth will be determined by the distribution of values for variables in whatever model (viz., system of variables-cum-structural equations) we are using, then we must first distinguish those variables in the model that depend (either immediately or mediately) on **X** from those that don't. In evaluating the given conditional, the latter variables have their values *held fixed* at whatever they actually are; only the values of the former are updated in accordance with the structural equations. The total set of values that results then determines the truth of **P**, and so the truth of the conditional. It is conditionals like this that, at least to a close approximation, capture the notion of what would happen as a result of an "intervention" on **X**. ("To a close approximation", because Woodward's official theory of interventions is a bit more complicated, for reasons that, though sound, we can afford to overlook.) Notice that this notion—so far, at least—appears to be *model-relative*. We'll come back to this point shortly.

Second, Woodward shows how to use systems of structural equations to define a number of interesting type-level causal relations among variables, as well as a token-level relation of "actual causation" between events (conceived of, roughly, as the instantiation of a particular valuable value by particular variable on a particular occasion; note that actual causation corresponds most closely to the relation that the mainstream literature on causation takes as its target of analysis). Thus, in the system depicted in figure 1, the variable **D** is a "direct cause" of variable **E**, while **C** is not a direct cause of **E**, but is a "total cause". In addition, in figure 1, the firing of neuron C counts as an actual cause of the firing of neuron E (translated, **C=1** is an actual cause of **E=1**).

Third, Woodward argues quite persuasively that the explanatory value of a given system of structural equations hinges crucially on the degree to which the relationships represented by these equations exhibit what he calls "invariance". Roughly, invariance consists in stability under various possible counterfactual perturbations of the system and its environment. It is of the first importance that these relationships remain stable under counterfactual variations in the values of the constituent variables; without this much invariance we have no explanatory power at all. But it is also important how stable they remain under perturbations of the environment; structural equations that describe dependency relations that are themselves too easily disrupted by alterations in the environment will be of little scientific interest. (Imagine that it was only in highly specific circumstances, difficult to produce and difficult to maintain, that the neurons in figure 1 would exhibit the relationships captured by the structural equations for that system; if so, we would naturally look for a much more encompassing system of variables plus structural equations—one much less susceptible to environmental disruption—in order to understand their behavior.) Woodward argues, again quite incisively, that it is the search for invariant relationships that drives scientific inquiry more than the search for "laws of nature".

Finally, Woodward offers a simple but compelling back story to address questions all too often overlooked in the philosophical literature on causation and explanation. Why is it that we come equipped with the causal and explanatory concepts we do? What is their value to us? Certain views on causation can look downright bizarre when confronted with these questions. For example, consider Lewis's counterfactual analysis of causation, together with his unwieldy prescription, based on the notion of "miracles", for evaluating counterfactuals. Woodward, building on criticisms of Horwich and others, makes what I think is a rather devastating case that our allegiance to this particular causal concept—that is, one understood in terms of the peculiar Lewisian recipe for evaluating counterfactuals—should seem wholly arbitrary. By contrast, Woodward offers up an account of our causal and explanatory concepts as having the structure they do because, fundamentally, we are agents with an interest in manipulating the world around us. It is not that these concepts are somehow subjective. No, they are concepts of perfectly objective dependency relationships that exist in the world, and in particular Woodward is quite clear that he has no interest in reducing the concept of causation to some concept of human agency. His point is more subtle: if we wish to understand why, among all the possible objective relationships we might have latched on to, these ones hold such pride of place in our conceptual repertoire, we can do so by seeing how conceiving of the world in terms of them contributes to our aims as agents with an interest in manipulating our surroundings.

So much by way of overview. Let me turn now to what I think are two of the more important criticisms of Woodward's approach—or perhaps better, to two of the more important bits of business left unfinished by what he offers us.

The first issue concerns the pair of foundational questions: what are variables? And what are the truth conditions for structural equations relating them? Now, while it is certainly the case that, in working through Woodward's book, one develops an excellent feel for how to answer these questions in particular cases, there's very little Woodward says that explicitly addresses them. (He is not alone in this. The literature on structural equations tends to be quite relaxed in its attitude towards these questions.) Woodward is perfectly forthright that he does not intend to give any sort of reductive analysis of structural equations, pointing out that the account he gives of them in terms of the notion of intervention is circular, in so far as the notion of intervention is itself a causal one, to be elucidated by appeal to structural equations themselves. This seems not to bother him, in part because the circle in question is far from unilluminating. For the interrelationships Woodward displays between the various bits of his apparatus unquestionably introduce substantive constraints. But in addition, Woodward is skeptical of the possibility of any kind of reductive account of causation, and more specifically of any kind of reductive account of the dependency relations that structural equations aim to capture.

But I think his skepticism is a bit hasty, and more importantly that there are good practical and theoretical reasons for pursuing these two foundational questions in depth. To see what they are, let's first explore what an answer to them might look like.

Neuron diagrams are easy to model, in large part because it is so easy to choose variables for them: For each neuron, and each relevant time-interval (roughly: each time interval such that that neuron could, in the circumstances, do something interesting in that time-interval), we introduce a variable to correspond to the state of that neuron during that interval. Typically, two values will suffice, one for "firing" and one for "not firing". But we can easily add values, if we wish to distinguish different *ways* the neurons can fire.

All of this suggests a more general prescription for choosing variables and values, for an arbitrary system we might wish to model: First, find a way to "carve up" the system into discrete, well-defined sub-systems. Second, for each relevant sub-system, and each relevant time or time-interval, introduce a variable to characterize the intrinsic physical state of that sub-system at that time, or during that time-interval.

I used "relevant" twice, partly because not every sub-system needs explicit representation (for example, one need not bother with variables corresponding to the stimulatory and inhibitory channels between neurons), and partly because not every

moment of time is such that the behavior of the system at that time needs representing (for example, one also need not bother with variables that characterize the state of neurons before or after the events under consideration, or during the passage of signals).

In addition, it may not always be straightforward how to "carve up" the given system into sub-systems. It will be *fairly* straightforward, if the system is constituted by a number of clearly distinguishable, interacting parts. But that won't always be the case—at least, at the desired level of description. Consider the flow of water down some rapids: what choice could we make of interacting parts, given that we don't wish to introduce variables for the state of each water molecule at each moment? Here a kind of default option suggests itself, which is that we choose variables to correspond to reasonably well-defined *regions of space* at different times, or regions of spacetime. The price of exercising this option is, in general, that no set of variables will stand out as uniquely appropriate.

Patently, what I've offered is far very from an exact recipe for determining the variables and values appropriate for modeling any given situation. But that is perhaps as it should be: within broad but non-trivial constraints, many choices are permissible. Still, my approach differs substantially from Woodward's own. Consider a sample passage:

> Intuitively, variables are properties or magnitudes that, as the name implies, are capable of taking more than one value.… Many of the familiar examples of so-called property causation discussed in the philosophical literature may be understood as relationships between two-valued or binary variables, with the variable in question taking one of two values, depending on whether the properties in question are instantiated or not. Thus, the claim that ingestion of aspirin causes recovery from headache may be understood as asserting a relationship between the values of a variable *A*, representing whether or not aspirin is ingested, and the values of a variable *H*, representing whether of not relief from headache occurs. (p. 39)

Variables, for Woodward, are thus *type-level* entities, whose job is *not*, in the first instance, to characterize the state of a localized bit of the world at a specific time. By contrast, I suggest that we think of variables precisely as vehicles for spatiotemporally localized representation of the world. An advantage of doing so will emerge shortly, in the form of a relatively clean account of what it is for a structural equation to relate some variable.

Okay, suppose we've chosen our variables for some system of interest. What are the truth-conditions for structural equations? *What is it* for such-and-such structural equations to correctly relate these variables? Woodward doesn't say, and it seems

clear that much of his tolerance for the lack of a clear, reductive account of the truth-conditions for structural equations comes from the impression that there are no adequate alternatives to treating the notion as a kind of primitive, and using *it* to analyze the counterfactuals that will provide the ingredients for an account of causation and explanation. Now, to the extent that one thinks of Lewis's "miracles"-based account as providing the main or perhaps only alternative, this impression is entirely understandable. Indeed, Woodward does a thorough and brilliant job of showing that that account is baroque, poorly motivated, and fails even on its own terms. But there is a much better account, which builds on a proposal of Tim Maudlin's. It will work *quite* nicely to provide truth-conditions for structural equations, provided the modest strictures on choice of variables discussed above are adhered to. To see how it works, and how it can be adapted to the needs of causal modeling, let's begin with a simple example.

At noon, Suzy throws a rock at a window. A few second later, the window breaks. Let C be her throw, E the breaking of the window. We seek truth-conditions for the conditional "if C had not occurred, then P", where P can be any claim about the post-C history of the world (e.g., "E does not occur"). So consider the state of the world—the complete, fundamental physical state of the world—at the time at which C occurs. Consider a nomologically possible alternative to this physical state, which is just like it except that C does not occur. (Think of arriving at this state as follows: Begin with the actual state. Make localized changes to it—localized, that is, to the place or physical systems involved in C's occurrence—sufficient to guarantee C's non-occurrence.) The actual, fundamental physical laws proscribe a certain forward evolution for this physical state. If that forward evolution is such as to make it the case that P, then the conditional is true; if it is such as to make it the case that not-P, then the conditional is false. In the given example, we begin with an alternative state that is just like the actual state, save that Suzy doesn't throw. Forward evolution: the window doesn't break. In this way we secure the intuitively correct value *true* for "if Suzy hadn't thrown, the window would not have broken".

We can also use this recipe to evaluate conditionals of the form "if C had occurred in such-and-such a manner, then P": Begin with the actual state of the world at the time of C's occurrence. Make localized changes, sufficient to make C occur in the specified way. Evolve the resulting state forward, in accordance with the actual fundamental laws. Check to see whether P comes true. Shortly, this version of the recipe will prove useful, in given systematic truth-conditions for structural equations.

A number of comments, before we proceed to that task.

First, these truth-conditions for counterfactuals don't yet take proper account of indeterminism, of either the fundamental stochastic variety or the statistical me-

chanical variety. Nor do they take account of the relativistic prohibition on talk of states-at-times. We'll ignore these issues here.

Second, it's worth emphasizing that this account breaks sharply with philosophical tradition, in that it does not give a semantics for counterfactuals in terms of similarity between *possible worlds* but rather in terms of similarity between possible *complete physical states* of worlds. A smart move: much of what is so baroque about Lewis's account, for example, flows from his insistence on defining a similarity relation between whole worlds. Note, in addition, that not much is required here of the notion of "similarity": what we need, ultimately, is just a well-defined sense in which one complete physical state can be exactly the same as another, except in a certain specified respect that concerns a localized region or physical system.

Third, we should not think that when we modify this localized region or physical system so as to make some actual event C fail to occur, we try to find an alternative state for this patch of the world that is as similar as possible to its actual state, consistent with the requirement that C not occur. That will lead to silly deliberations like the following: "Well, in the counterfactual situation in which Suzy's throw does not occur, what happens instead? Does she perhaps toss it? But then how do we know that such a tossing is not numerically identical to the *actual* throw?" A much better view is that for any given event, we work with an antecedently understood distinction between a *default state* for the region in which the event occurs, or for the physical system or systems to which it pertains. Conceiving of the event as one among various possible *deviations* from that default state, we answer the question, "What would have happened, had that event not occurred?" by returning the relevant region or system to its default state, holding the state of everything else fixed. In the case of Suzy, what we naturally think is that if she had not thrown the rock, what she would have been doing *instead* is standing there idly—*doing nothing*, as it were. Likewise, if we ask what would have happened, had a given neuron-firing not occurred, we naturally focus on an alternative situation in which that neuron remains, at the time in question, *in its dormant state*—not a situation in which it fires in a different manner.

Fourth, it is not really to be hoped that—even with a default state specified—the recipe will yield a *unique* counterfactual state of the world. Here, multiple realization reigns, and we should correspondingly expect limits on what we can say about any given counterfactual situation. If Suzy's throw hadn't occurred, the window wouldn't have broken. –Not just *then*, at least. Would it have *remained* unbroken for the next year? We don't know, of course, and not because it's too hard to find out!

Fifth, the recipe is quite limited in scope. It says nothing about conditionals such as the following: "If gravity had obeyed an inverse-cube law, Kepler's second law still would have held." (True, by the way.) Nor is it built to handle "backwards" condi-

tionals, in which the consequent concerns a time or times *before* the time or times that the antecedent is about. Neither limitation poses a problem, given the purposes to which we will put the recipe; the second, in fact, is exactly what allows us to avoid talk of "miracles" (for we simply don't care how the world managed to get into the counterfactual state specified in the antecedent).

Time to consider how to arrive at structural equations. We'll start with a simple idea, spot the need for an amendment, and refine accordingly.

As an illustration, suppose we have some situation for which we wish to provide a causal model, and suppose we've decided that this model should make use of five variables: $\mathbf{C_1}$, $\mathbf{C_2}$, $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{E}$. Respecting the strictures laid out above, we have chosen these variables in such a way that each has a well-defined time or time-interval associated with it. Let's suppose that $\mathbf{C_1}$ and $\mathbf{C_2}$ concern the same time, as do $\mathbf{D_1}$ and $\mathbf{D_2}$; let's suppose further that the temporal order among the five variables is this: $\mathbf{C_1}$, $\mathbf{C_2} < \mathbf{D_1}$, $\mathbf{D_2} < \mathbf{E}$. Then a simple approach is to stipulate that the equations for $\mathbf{D_1}$ and $\mathbf{D_2}$ shall include only $\mathbf{C_1}$ and $\mathbf{C_2}$, and that the equation for $\mathbf{E}$ shall include only $\mathbf{D_1}$ and $\mathbf{D_2}$. The Maudlin-recipe applies straightforwardly. To fix an equation for $\mathbf{D_1}$, for example, we need to determine, for each setting of the $\mathbf{C}$-variables $\mathbf{C_1 = v_1}$ and $\mathbf{C_2 = v_2}$, a resulting value for $\mathbf{D_1}$. Begin with the state of the world at the time the $\mathbf{C}$-variables concern. Modify it locally so as to make $\mathbf{C_1 = v_1}$ and $\mathbf{C_2 = v_2}$. Evolve the resulting state forward in accordance with the actual fundamental laws. The value for $\mathbf{D_1}$ will be that unique value $\mathbf{w}$ such that the proposition $\mathbf{D_1 = w}$ is guaranteed to be true, given this forward evolution. Don't be fooled, of course, by the heuristic talk of "modifying" and "evolving" (as if complete physical states were something we could manipulate). When cleansed of such talk, it's apparent that what we have provided here is a purely *metaphysical* story about what makes a given structural equation correct.

(What if there is no such unique value? Well, there won't be *more* than one; the worry is that there might not be *any*. If so, that shows that there was something wrong with our choice of variables—e.g., $\mathbf{C_1}$ and/or $\mathbf{C_2}$ weren't "fine-grained" enough, or $\mathbf{D_1}$ and/or $\mathbf{D_2}$ were *too* fine-grained. Then we should simply fix the problem, and move on.)

This account of the truth-conditions for structural equations almost works. But there is a problem, one that arises if we are, as it were, too parsimonious in our choice of variables. Consider figure 1 again. Suppose we simply *omit* the variables $\mathbf{A}$ and $\mathbf{B}$, choosing to construct a model using only $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$. Then the account just given will yield these structural equations:

E = D

D = C

There's no problem with the second, nor—in a *certain* sense—with the first; that is, the first correctly captures the way that **E** immediately depends on **D**. But put them together in a single causal model, and that model will tell us that the conditional

if C = 0, then E = 0

is *true*. And that is because, *according to the model*, an intervention on **C** that sets its value to 0 will have the effect (given the second equation) of setting the value of **D** to 0, which will in turn have the effect (given the first equation) of setting the value of **E** to 0. So the model fails to return the truth-value for the conditional "if **C** = 0, then **E** = 0" that we evaluate using the Maudlin-recipe. The latter conditional is straightforwardly *false*—false *full stop*, and not merely relative to this or that model. Perhaps we should rest content with the position that the former conditional can be true relative to some models (e.g., this one), and false relative to others (e.g., the more complete model of figure 1 provided earlier). Woodward does not, as far as I can tell, explicitly endorse this position, though other authors have (e.g., Halpern & Pearl).

But that would be a mistake, a move that would shift too much of the burden of providing an adequate structural equations account of causation and explanation onto the project of producing the as-yet unwritten rules for choosing "appropriate" causal models. It's much better to lay down rules that guarantee a certain kind of *stability* in our causal models, so that a conditional like the foregoing one will receive the same truth-value relative to every model that assigns it one. There is a natural way to achieve this effect, one with the added benefit of guaranteeing that this truth-value will *match* the one yielded by the Maudlin-recipe.

Return to figure 1, and our overly parsimonious model for it that used only variables **C**, **D**, and **E**. The trouble we got into with this model derived from the fact that, in the counterfactual situation in which **C** has the value 0, one *consequence* of its having this value is that neuron B *fires*, which in turn guarantees that E fires. But our paired-down model contains no variable whose value could reflect the fact that B fires. One bad solution to this problem is to insist that an acceptable model contain a comprehensive enough set of variables, so that any relevant consequence of one variable's having a given value gets explicit representation in the values of other variables. I think that places too high a demand on the causal modeler, and at any rate there is a cleaner approach. To illustrate, I'll stick to this three-variable model for figure 1.

The temporal order of the variables is **C** < **D** < **E**. In writing down equations for these variables, we adopt the policy that for a given variable, *any* temporally prior variable is allowed to figure in its structural equation. Since **C** is the sole variable prior to **D**, we recover, using the Maudlin-recipe, the same equation for **D** as before:

D = C

But **E** is now allowed to functionally depend on both **C** and **D**. That means that, for each of the four ways of assigning values to **C** and **D**, we need to determine a resulting value for **E**. So consider the case **C = x** and **D = y**. Focus on the state of the world at the time that **C** concerns. Make local changes, sufficient to guarantee that **C = x**. (If **C = x** in actuality, no changes will be necessary.) Evolve the resulting state forward *until the time that* ***D*** *concerns*. Make *local* changes to *this* state, sufficient to guarantee that **D = y**. (Again, no changes may be necessary.) Evolve this newly modified state forward in time. Some value for **E** will result. That is the value that the structural equation for **E** should specify as output, when given as input the values **C = x**, **D = y**.

Let's test this approach. For the situation depicted in figure 1, the actual values **C = 1**, **D = 1** obviously map to **E = 1**. Given the values **C = 1**, **D = 0**, we begin with the (actual) state in which both C and A fire, evolve forward into a state in which B doesn't fire but D does, *locally modify* this state so that D *does not* fire (and B still doesn't), evolve *this* state forward, and see that E does not fire. So, for **C = 1**, **D = 0**, we must have **E = 0**. It's routine to check the other two cases: **C = 0**, **D = 1** gives us **E = 1**; **C = 0**, **D = 0** gives us **E = 1**. More simply:

E = 1 – C + CD

Notice, finally, that the fact that this is the correct equation for **E** depends crucially on what variables are included in the model. Reintroduce **B**, for example, and the correct equation renders **C** irrelevant. That result, of course, is exactly as it should be.

The generalization of this recipe is straightforward: Suppose we have some variable **X** in some model M. If we have been scrupulous in our choice of variables, there will be a clear-cut distinction between those other variables in M that are temporally prior to **X**, and those that are not. For each way of assigning values to the former variables, we can follow the 'sequential updating' variant of the Maudlin-recipe to fix a resulting value for **X**. In this way, the fundamental laws, together with

the actual history of the world, will fix a unique structural equation for **X** (in terms of the other variables in M).

Now consider some variable **C** in M that is temporally prior to **X**. Consider the counterfactual situation in which **C** = **c**, arrived at by locally modifying the state of the world at the time that **C** concerns so as to make **C** = **c**, and evolving this state forward in time. This forward evolution will yield some assignment of values to all variables in the model: those that are temporally prior or concurrent with **C** will receive their *actual* values; the remaining variables may receive different values. What's more, given our truth-conditions for structural equations, the value that **X** receives in this counterfactual situation *must be the same* as the value that the structural equation for **X** yields, when given as input the values that all the variables prior to **X** receive, in this counterfactual situation. It follows that the conditional

if C = c, then X = x,

evaluated by Woodward's interventionist procedure (set **C** equal to **c**; update values of other variables in accordance with the model's structural equations), must, regardless of the details of the model M, receive the same truth-value as the counterfactual "if **C** = **c**, then **X** = **x**", when evaluated by the Maudlin-recipe. We have thus arrived at truth-conditions for structural equations that are not only clear, but that also guarantee that what may seem to be *model-relative* truth-conditions for conditionals are in fact *not* model relative: Any model that assigns a conditional a truth-value will assign it the same truth-value, and moreover will assign it the truth-value it *ought* to have (i.e., the truth-value determined by the Maudlin-recipe).

We now have reasonably good answers to the two questions raised at the outset: What are variables? What are the truth-conditions for structural equations? If something like the view I have been recommending about variables and structural equations is correct, then it has an immediate theoretical payoff, for it shows that Woodward's conception of explanation, far from being inimical to—or even an *alternative* to—the view that laws of nature play an important role in explanation, in fact *supports* that view. For explanation consists in exhibiting invariant relationships of the sort that structural equations aim to capture, and laws of nature—understood, now, as the sorts of things that fundamental physics and fundamental physics *alone* aims to uncover—provide an essential ontological grounding for these invariant relationships. The search for explanations is part and parcel of the search for laws—or perhaps better, is part and parcel of the exploration of the nomological structure of the world, a structure it has in virtue of its fundamental laws.

Having said this, Woodward's trenchant criticisms of the still-widespread understanding of *how* laws figure in explanations—namely, that they must somehow

"cover" the things to be explained in something like the way that the logical empiricist deductive-nomological model described—still stand, as do his remarks that the special sciences, at least, have invariant relationships and not laws as their most immediate target. On the view I'm suggesting, the only sorts of laws there are are fundamental laws, and their role in explanation, although absolutely essential, is unquestionably indirect.

Finally, the view I have sketched about variables and structural equations—as vague as it no doubt is in many respects—at least adds enough clarity to yield some practical benefit, in the form of cautionary remarks about how to model specific situations. For I worry that Woodward's discussion seriously underplays the mistakes one can easily fall into when constructing a model even for a very simple situation, and that avoiding these mistakes requires extremely close attention to the content of a particular assignment of variables, and of proposed equations relating them. To illustrate this point, I will rely on a classic example from the causation literature:

Suzy First: Suzy, an expert rock-thrower with a taste for minor acts of destruction, throws a rock at a bottle. The rock hits the bottle, shattering it. Suzy's friend Billy throws a rock at the bottle, too. He's just as expert as she is, but a bit slower. Consequently, her rock gets there first; but if she hadn't thrown it, the bottle would have shattered all the same, thanks to his throw.

It is supposed to be a major achievement of structural equations approaches that they offer a powerful new technique for dealing with such a stubborn counterexample to so many counterfactual analyses of causation. In fact, they fail rather miserably, and at a surprising stage: the causal models standardly offered as representations of cases like Suzy First are simply *incorrect*. I'll consider one example, taken from Halpern & Pearl (2005); while Woodward does not explicitly consider this case, the analysis that follows is exactly what one would expect, given his preferred structural equations treatment of actual causation.

Superficially, the model is quite elegant and simple; judging from various conversations I've had, it is currently thought of as providing the canonical structural equations treatment of Suzy First. It makes use of just five variables:

ST: has value 0 if Suzy does not throw; 1 if she does.

BT: has value 0 if Billy does not throw; 1 if she does.

SH: has value 0 if Suzy's rock does not hit the bottle; 1 if it does.

BH: has value 0 if Billy's rock does not hit the bottle; 1 if it does.

BS: has value 0 if the bottle does not shatter; 1 if it does.

We should understand each of these variables as making implicit reference to a particular time. More specifically, let's stipulate that Suzy throws at time 0, Billy throws

at (slightly later) time 1, Suzy's rock strikes the bottle at time 2, Billy's rock *would* have struck the bottle at time 3 (i.e., if Suzy's had not already done so), and the bottle is in a shattered state at time 4. So **ST** characterizes what Suzy is doing at time 0; **BT** what Billy is doing at time 1; **SH** what Suzy's rock is doing at time 2; **BH** what Billy's rock is doing at time 3; and **BS** the state of the bottle at time 4. What we wish to write down are structural equations that show that it is Suzy's throw, and not Billy's, that causes the bottle to be in a shattered state at time 4.

This seems to be easy to do. Halpern and Pearl—and just about everyone else, as far as I can tell—find the following equations satisfactory:

BS = BH + SH – BH•SH
BH = BT(1 – SH)
SH = ST

Assuming these equations are correct, Woodward's account of actual causation will judge Suzy's throw to be a cause of the shattered state, by virtue of the conditional

if (ST = 0 & BH = 0), then BS = 0

For if imagine an intervention that sets **ST** to 0 and **BH** to 0, then the third equation tells us that **SH** will be 0; the second equation becomes irrelevant (since the intervention breaks the connection between **BH**, on the one hand, and **BT** and **SH**, on the other); and the first equation tells us that **BS** will be 0. It's also not too hard to confirm that his account will judge Billy's throw *not* to be a cause of the shattered state (I'll omit the details).

Success? Not so fast. Let's take a close look at what these equations mean. The third is unobjectionable: it says that Suzy's rock will hit the bottle iff she throws it. Given that we only mean to be considering four options for the temporally prior variables **BT** and **ST**—she throws/doesn't throw at just the time and with just the speed she does; he throws/doesn't throw at just the time and with just the speed he does—this equation is perfectly correct. The first equation might also seem correct: for it says merely that the bottle will be in a shattered state iff at least one of the rocks hits it. Likewise the second, which says that Billy's rock will hit the bottle iff he throws it, *and* Suzy's rock hasn't already hit it (for in that case, it won't be there for his rock to hit).

But now we should smell a rat. Look again, and closely, at the first two equations. The first strikes us as true in part because, when we envision a situation in which **BH** = 0 and **SH** = 0, we understand that **BH** = 0 *because Billy's rock isn't thrown*, and instead lies idle (we may suppose) in his hand. But the second strikes us as true in part because, when we envision a situation in which **BT** = 1 and **SH** = 1,

we understand that **BH** = 0 *because the bottle isn't there to be hit*. The model is simply trading on this ambiguity in the content of the claim "**BH** = 0". Remove the ambiguity, and one or the other of the first two equations must be revised.

Let's work through this again, slowly and systematically, making explicit use of the truth-conditions for structural equations spelled out above. Those truth-conditions will straightforwardly vindicate the third equation. As for the second, we begin by observing that the variables that are candidates for figuring in the equation for **BH** are **ST**, **BT**, and **SH**, since these are the only variables temporally prior to **BH**. There are eight settings for these three variables that we need to consider. Following the sequential updating version of the Maudlin-recipe, we can immediately see that in any counterfactual situation in which **ST** = 1 and **SH** = 1, the bottle must be in a shattered state immediately after time 2 (so: before time 3). Forward evolution in accordance with the laws will give us a time-3 state of the world in which the bottle is *still* shattered; hence **BH** = 0, regardless of the value of **BT**—*provided* we understand "**BH** = 0" as meaning simply that Billy's rock fails to strike the bottle. (Shortly, we'll see reasons to understand it differently.) That takes care of two of the eight cases. Suppose next that **ST** = 0 and **SH** = 0. Then, clearly, **BH** = 1 iff **BT** = 1. That takes care of two more cases. There are two more cases in which **BT** = 0, in both of which **ST** ≠ **SH**; it's not clear what goes on with Suzy's rock in those cases, but at any rate we can be sure that **BH** = 0, since Billy's rock isn't even thrown. Now to the two remaining cases:

BT = 1, **ST** = 1, **SH** = 0. Here, the time-0 state of the world is just the actual state, and no local modifications are necessary until we reach time 2, at which point we need to adjust the state so that **SH** = 0. Now, just how exactly do we do this? What sort of "intervention" do we have in mind? If the claim "**SH** = 1" is supposed to mean that Suzy's rock strikes the bottle in a certain way—namely, the way in which it *actually* strikes the bottle—and if "**SH** = 0" is supposed to be true iff "**SH** = 1" is false, then the problem is that there are far too many ways to locally modify the state so that **SH** = 0, and no principled way to choose among them. Worse: *some* of these modifications will yield forward evolutions in which the bottle is shattered, in which case we won't get the result that, for this choice of values, the correct equation must yield **BH** = 1. For example, Suzy's rock might strike the bottle in a rather different way from how it actually does, but still hard enough to break it.

Well, can't we just read "**SH** = 0" as saying that *Suzy's rock doesn't strike the bottle*? If so, it obviously doesn't strike it in a different way! But that isn't really any help, for we are still left with the mystery as to how to locally modify the state of the world, so

as to secure the truth of this claim. More to the point, *one* way to effect this modification is to have the bottle *be in a shattered state* before Suzy's rock can strike it, whence forward evolution will give us the unwanted **BH** = 0, as before. How, exactly, are we supposed to rule out such a modification as illegitimate?

As far as I can tell, the only clean, non-ad-hoc way to secure the desired result is to read "**SH** = 0" as meaning that Suzy's rock is simply *absent* (absent, that is, from the neighborhood of the bottle—perhaps we should return it to Suzy's hand…), at the relevant time. Let us so read it. Then granted: **BH** = 1.

BT = 1, **ST** = 0, **SH** = 1. Here, the time-0 state of the world is one in which Suzy is not throwing, but, as in the actual world, Billy is preparing to throw. Forward evolve this state until time 1. Billy throws (so no local modifications to the state are necessary). Forward evolve the resulting state until time 2. Make local modifications, so that Suzy's rock—which, remember, has been sitting in her hand—hits the bottle. Once again it's not so clear how to proceed, since we haven't said with enough specificity what the content of the claim "**SH** = 1" is. So let's correct that oversight, by stipulating that this claim says that Suzy's rock hits the bottle in just the way it does in the actual situation. Now we can proceed: the bottle breaks. Forward evolving, we see that Billy's rock doesn't strike the bottle. But it doesn't follow that **BH** = 0—*that* depends, unsurprisingly, on what the precise content of this claim is. If we take our cue from the foregoing discussion of "**SH** = 0", we will say that "**BH** = 0" is *false* in this situation, since Billy's rock is *not* absent from the given region: it's there all right, it's just flying over scattered shards of bottle-glass. But presumably this is not what Halpern and Pearl have in mind. So let's take it that "**BH** = 0" means simply that Billy's rock doesn't strike the bottle. Then granted: **BH** = 0.

We've now secured the second of the three structural equations, albeit at some cost: we were forced, somewhat surprisingly, to treat "**BH** = 0" as not at all analogous to "**SH** = 0". As you may have guessed, there is worse trouble ahead. We run into it as soon as we try to write down an equation for **BS**.

Consider the values **BT** = 1, **ST** = 0, **SH** = 0, **BH** = 0. What should be the corresponding value for **BS**? It should be **BS** = 0, we are told—after all, that we are describing a situation in which neither rock strikes the bottle. But having been alerted by the foregoing discussion, we can easily see how this response involves some sleight-of-hand. Let's work through the case systematically, to pin down where the fallacy is lurking.

For the given values of the 'input' variables, we start with a time-0 state locally modified from the actual state, so that Suzy does not throw. Evolve forward to time

1. Billy throws (no modification necessary). Evolve forward to time 2. Suzy's rock is absent from the region of the bottle (no modification necessary). Evolve forward to time 3. Billy's rock is about to strike the bottle—and now we need to make local adjustments, so that it doesn't. Not *not* NOT so that his rock is *absent*: we know how to do that (just change the world-state so that no rock is there, replacing it by air), and we know that forward evolution of such a modified state will yield a time-4 state in which the bottle is not shattered. But all this is entirely irrelevant, since we already know, from having thought through the equation for **BH**, that "**BH** = 0" had better *not* mean that Billy's rock is absent from the region of the bottle (else that equation is simply *incorrect*); rather, it had better mean only that Billy's rock—one way or another—does not strike the bottle. And with that meaning, the instructions to locally modify the state of the world so that **BH** = 0 are simply too ambiguous: *one* way is to remove Billy's rock, but *another* way is to change the state of the bottle from whole to shattered. These different modifications yield different forward evolutions—and different values for **BS**. So we cannot, after all, even *write down* a correct equation for **BS**.

What is the root of the problem? The key is to recognize that at time 3, there are *three* distinct states of affairs that we will want the variable **BH** to represent. **BH** should have one value if Billy's rock is simply absent from the region of the bottle (which is what will happen, if Billy doesn't throw). It should have another value if Billy's rock is striking the bottle (which is what will happen, if Billy throws but Suzy doesn't). And it should have a third value if Billy's rock is flying over scattered shards of bottle-glass (which is what will happen, if Billy and Suzy both throw). Let these values be 0, 1, and 2, respectively. As for **SH**, we will keep it two-valued: **SH** = 0 if Suzy's rock is absent from the region of the bottle; **SH** = 1 if it strikes the bottle. *Now* we can apply the Maudlin-recipe cleanly; for we no longer are trying to make the single value **BH** = 0 do double duty, signifying that Billy's rock is absent, when used in the equation for **BS**, but that Billy's rock doesn't strike the bottle, when derived using the equation for **BH**. The following equations result:

BS = sign(SH + BH)
BH = (SH + BT)BT
SH = ST

Now that we have a causal model that we need not feel ashamed of, do we at last get the right result—that Suzy's throw, and not Billy's, is a cause of the bottle's shattered state? That depends, of course on how one wishes to exploit structural equations in giving an account of causation. But it won't be straightforward. One of the

simplest approaches (the preliminary account favored by Woodward) is to look for a path from **ST** to **BS**, and ask whether **BS** depends on **ST**, if the values of variables *off* this path are held fixed. The only relevant path from **ST** to **BS** is **ST-SH-BS**. The off-path variables have values **BT** = 1 and **BH** = 2. Either of two conditionals could thus testify to **ST**'s causal status with respect to **BS**. But both of them are false:

if (ST = 0 & BT = 1), then BS = 0

if (ST = 0 & BH = 2), then BS = 0

Suppose we had reported the actual value of **BH** this way: **BH** ≠ 1. Then we might think to confirm the causal standing of **ST** by means of the following conditional:

if (ST = 0 & BH ≠ 1), then BS = 0

This conditional can seem to be *true*, if what we have in mind as a situation in which **BH** ≠ 1 is a situation in which the bottle remains intact, but Billy's rock somehow never reaches it. But, as soon as we take care to distinguish this situation from the situation in which Billy's rock does not strike the bottle *because the bottle is already shattered*, we expose this reasoning as fallacious. In fact, the correct causal model simply fails to assign a truth-value to this conditional. And that is because the antecedent is ambiguous. Disambiguated one way, we get the relevant but *false* conditional

if (ST = 0 & BH = 2), then BS = 0

Disambiguated the other way, we get the true but *irrelevant* conditional

if (ST = 0 & BH = 0), then BS = 0

There are, of course, other ways one might try to secure the intuitively correct judgments about the causal structure of Suzy First, within a structural-equations approach. Suffice it to say that none that I know of is without serious problems. At any rate, the main lesson I wish to drive home is different: the cavalier and sloppy treatment of this kind of example within the structural equations literature seems to me to stem directly from the relaxed attitude that literature displays towards our two foundational questions. It is business that really needs to be attended to.

The second bit of unfinished business alluded to above concerns the nature of the facts relevant to our causal and explanatory judgments. Woodward takes them to be exhausted by the sorts of facts captured by an adequate system of structural equations, but I'm not so sure, for two reasons. The first of these hearkens back to Suzy First: it seems to me that we ground our judgment that Suzy's throw is a cause of the shattering and Billy's is not in, roughly, the observation that the *intrinsic character of the process connecting her throw to the shattering* is of the right sort. If this is correct, it suggests

that an adequate analysis of the causal structure of at least some situations must attend to more than relations of counterfactual dependence. (See Hall 2004.)

The second reason can be brought out by comparing the events depicted in figure 1 with those depicted in figure 3, a slight variant on figure 2.

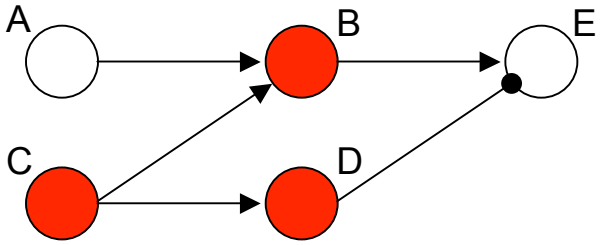

Figure 3

A natural, intuitive gloss on figure 1 is this: the firing of C causes the firing of E, and, at the same time, prevents A from being a cause of the firing of E. But the intuitive causal structure of figure 3 seems quite different: in particular, it does not seem right to say that the firing of C prevents the firing of E—let alone that it stops the non-firing of A from preventing the firing of E! Notice, however, that if we look just at the abstract patterns of counterfactual dependence, they are exactly the same as between figure 1 and figure 3. (This may be obvious; if not, take it as an exercise.) What would account, then, for the apparent difference in causal structure? Some authors (Hitchcock, Maudlin, and myself, for example) have recently speculated that one thing that matters to us is which values for variables count as *default* values—very roughly, the value a variable will have if nothing interferes with the system in question. (Remember that in the discussion, above, of certain kinds of counterfactuals, we already saw the need to invoke some such distinction.) Now, like Woodward, I am wary about leaning too heavily on intuitions about cases. In fact, his attitude, expressed quite elegantly in the following passage, seems to be exactly right:

> My interest in this section has been in showing how the apparatus of directed graphs and a manipulationist approach to causation can be used to reconstruct commonsense judgments about token-causal relationships. I want to conclude, however, on a somewhat more skeptical note. If the discussion in this section has been successful, what it has accomplished is [to] successfully isolate facts about patterns of counterfactual dependence, as revealed in hypothetical manipulations, that are relevant to commonsense token-causal judgments and causal distinctions. However, in at least some of the cases discussed above, it is controversial what the deliverances of common sense are and even more so whether (or even what it would mean to say that) such deliverances are "correct". The suggestion I want to make is that to the extent that commonsense causal judgments are unclear, equivocal, or disputed, it is better to focus directly on the patterns of counterfactual de-

> pendence that lie behind them—the patterns of counterfactual dependence are, as it were, the "objective core" that lies behind our particular causal judgments, and it is such patterns that are the real objects of scientific and practical interest. (p. 85)

I take the lesson to be this: if we agree on the sorts of facts that our causal talk is aiming to get at, and agree on how, in a particular case, those facts are arranged, then any lingering disagreements about what causes what should be viewed as terminological. In the present case, if we agree that the facts that matter are just the facts about abstract patterns of counterfactual dependence captured by structural equations, then we should view any disagreement about how to describe figures 1 and 3 as a terminological quibble, and insist that in the only causally important respects, the structures depicted are exactly the same. But the fact that intuition recognizes a difference between these cases might lead us to a different verdict, which is that our causal talk aims to latch onto a richer structure than that given merely by the abstract patterns of counterfactual dependence. I see no deep threat here to a structural equations approach, especially since a distinction between default and deviant values can quite easily be accommodated within that approach. But I do think we should take it as a question for ongoing research what, over and above the facts captured by structural equations, might matter to us when we represent the world in causal terms. One of the many virtues of Woodward's book is that it allows philosophically rich questions such as these to be posed in such sharp terms.

References

Hall, Ned 2004a: "The Intrinsic Character of Causation," in Dean Zimmerman (ed.), *Oxford Studies in Metaphysics, Volume 1*:255-300.

Halpern, Joseph Y., and Pearl, Judea 2005: "Causes and Explanations: A Structural-Model Approach. Part 1: Causes", *British Journal for the Philosophy of Science* 56: 843-887.